Analysing and Reporting Performance Indicator Data: 'Caress' the data and user beware!

Ken Rowe
Australian Council for Educational Research

Background paper[1] to invited address presented at the
2004 Public Sector Performance & Reporting Conference,
under the auspices of the International Institute for Research (IIR)
Sydney, 19-22 April 2004

Abstract: Within the context of a growing international movement toward the adoption of 'outcomes-based' modes of public sector service provision, policy, governance and accountability, this paper focuses on the context, nature and purpose of performance indicators (PIs), and in particular on the analysis and reporting of data derived from them as bases for informing policy. Presented and discussed are key issues related to:

- The *nature*, *purpose*, *types* and *sources* of PIs;
- essential features of *useful* indicators;
- the 'dangers' of analysing, interpreting and reporting **aggregated** data, and
- effective methodologies for the responsible analysis and reporting of PI data.

To illustrate these 'issues', specific reference is made to **educational PIs**, and especially to responsible modes of PI data analysis, use and reporting.

1.0 What are performance indicators?

In general, *performance indicators* (PIs) are defined as *data indices of information by which the functional quality of institutional service providers and systems may be measured and evaluated*. Typically, within the context of specified 'target objectives', PI data are 'measures' of operational and functional aspects of organizations and/or systems that provide evidential bases for determining the extent to which such 'target objectives' are being met. PIs serve various purposes, the most notable of which are for **monitoring**, **policy determination**, **target-setting, evaluating** and **reforming**. Although the essential features of PIs for public sector institutions are consistent with their counterparts in private sector corporate enterprises, they also have unique characteristics – key aspects of which are highlighted and illustrated in this paper, with specific reference to educational PIs.[2] At the outset, however, it is helpful to note the importance of educational PIs in prevailing local and international contexts.

1.1 The nature, purpose, types and sources of educational PIs

During the past twenty five years, education systems throughout the world have been subject to considerable reform and change – all justified on the grounds (or at least the rhetoric) of **improving** the ***quality*** of school education. A key feature of this change has been the frequent revisions of style and policy focus, especially in the area of PIs, with major emphases being placed on the assessment and monitoring of student learning outcomes. Indeed, current policy activities related to 'outcomes-based' educational PIs and their links with growing demands for

---

[1] Correspondence related to this paper should be directed to: Dr Ken Rowe, Research Director (Learning Processes & Contexts), Australian Council for Educational Research, 19 Prospect Hill Road (Private Bag 55), Camberwell, Victoria 3124, Australia; Tel: +61 3 9277 5584; Fax: +61 3 9277 5500; Email: rowek@acer.edu.au; Web: www.acer.edu.au.

[2] See: Rowe (2000a,b, 2001); Rowe and Cresswell (2002); Rowe and Lievesley (2002); Rowe, Turner and Lane (2002); Visscher and Coe (2002).

accountability, *standards monitoring*, *benchmarking, school effectiveness* and *reform* are widespread and well established in many developed countries.[3]

Such emphases are aptly illustrated in the reported proceedings of a meeting under the auspices of the *Summit of the Americas* (2002), which states:

> Although it is now part of daily life in schools and in debates between specialists, education assessment has recently become a relevant topic for governments and society, especially because of the economic crisis and the acceleration of the globalization process, which make investments in education a strategic point while the resources available for the sector have shrunk.

In many developed countries, including Australia, much of this activity has been (and continues to be) directed away from concerns about *inputs* and *processes* of educational systems (e.g., physical resources and curriculum provision) to *outputs* (e.g., improvements in student achievement outcomes, as well as in school and system performance). A major effect of such activity has been to signal shifts in government policy intention to: (a) encourage system accountability to ensure both efficient and effective utilization of resources, and (b) bring the delivery of educational services into public sector accounting, underscored by a concern to ensure that such services represent 'value for money'.

Since schooling accounts for significant proportions of both public and private expenditure, as well as generating a substantial quantity of paid employment for teachers and administrators, the enduring interest by governments (and their supporting bureaucracies) in the relative *performance* of education provision is not surprising, particularly in primary and secondary schooling. This is an especially sensitive issue at the present time given the level of consensus regarding the importance of school education as an element of micro-economic reform and in meeting the constantly changing demands of the modern workplace – within increasing world economy globalisation trends (Mortimore, 2001; NCEE, 1997). Proclamations by the international media magnate Rupert Murdoch at the National Press Club on 12 October 2001, serve to underscore this importance for Australia's current and future economic viability. On this occasion, Murdoch asserted that "…if Australia continues with its reluctance to invest in the quality of its primary, secondary and tertiary educational infrastructure, Australia will end up even further behind the international economic '8-ball' than it is at present, such that Paul Keating's 'banana republic' prognostications will become a reality".

Whereas the provision of quality education is critical to the development of all countries, it is especially the case for developing countries where there is considerable pressure to increase access to education, but not at the expense of quality. Hence, the demand is to ensure that PIs do not provide a partial, and thus potentially misleading picture of either *quality* or *effectiveness*, as has often been the case in the past. Despite the difficulties entailed in defining *educational effectiveness* at the school, system or national levels, and reaching consensus on the relevant criteria, a good deal of discussion has focused on what is meant by *quality schooling*, and how it might be measured and improved. Although the term *quality* is likewise problematic, the "...measurement of the *quality* of schooling is of critical importance at a time when so much school reform in so many parts of the world is being undertaken." (Mortimore, 1991, p. 214) In fact, concerns about the *quality* of school education and its monitoring have long been high priority policy issues in all OECD countries (see: OECD, 1983, 1986, 1989, 1993, 1995). To date, the major indicators of *educational effectiveness* from which judgments of *quality* are derived, continue to focus on measures of student achievement outcomes – particularly in Literacy, Numeracy and Science. In this context, Manno (1994) has noted:

> When judging educational quality, either we focus on what schools spend – or one of its many variants – or we focus on what students achieve, what they know and can do. Those who advocate a focus on outcomes in judging educational quality hold one common belief: we must specify what we expect all children to learn, and we must assess them to determine whether they have learned it.

---

[3] See especially: Alton-Lee (2003); Buckingham (2003); Dorn (1998); Hill and Crévola (1999); Monk (1992); Rowe (2003); Tucker and Codding (1998); Visscher *et al*. (2000).

Although measures of student learning outcomes are prime PIs of education systems and the services they provide and for which they are responsible, there are many others (including *inputs, processes* and *outputs* – including students' attitudes, values, behaviours and related social outcomes of schooling) that constitute useful bases for informed planning and decision-making, followed by implementation and reform. If decisions for improvement are to be **informed** rather than based on political/bureaucratic whim or ideology, useful, dependable and timely information on indicators is required. Indeed, such bases constitute key purposes of specifying, gathering and using PIs for educational change and reform. In particular, PI information allows systems (educational and otherwise) and their constituent organizational elements to: (1) formulate strategic policy priorities and their related targets, (2) specify achievable objectives, (3) implement them, and (4) evaluate the extent to which those target objectives have been attained.

1.2 Types and sources of educational PIs

The types of input-output PIs are many and varied. Among the major educational PIs that may be collected include:

- Indicators of resource provision and funding, specified against stipulated targets;
- Access rates at: pre-school, primary, secondary, vocational and tertiary levels – per capita of age/stage cohort population, and inequities in access to education;
- Participation rates in education at all levels, barriers to participation; characteristics of children and adolescents out of school;
- Repetition rates and completion of five to twelve years of schooling;
- Percentage of GDP devoted to education;
- Per capita costs at each of these levels;
- Class sizes; teacher : student ratios;
- Provision and up-take of teacher education, training and participation in in-service professional development;
- Measures of student achievement outcomes in core curricular at specified age/grade-levels or cohorts;
- Longitudinal achievement progress indicators and measures of factors affecting students' progress rates;
- Measures of impact of strategic interventions for students with special needs and those from disadvantaged backgrounds; and
- 'Value-added' indices of measured outcomes and service provision (i.e., net-effects on progress in excess of that predicted from initial outcomes and context measures).

Sources from which educational PI data may be obtained are inherently multilevel and multi-faceted. That is, data can be gathered from multiple levels of a system, namely: student, class, school, district, region or province, state, national and international (e.g., such as the *Third International Maths and Science Study* (TIMSS)[4] and the *OECD Programme for International Student Assessment* (Lokan, Greenwood & Cresswell, 2001; OECD-PISA, 2001). PI data may also be gathered from administrative records, school surveys, household surveys and population censuses. The ways such data are gathered range from rudimentary manual methods to sophisticated computer-based management information systems implemented by governments and their supporting bureaucracies (see below). In the latter case, the rapid development of information and communication technology, increased pressures to collect and 'measure' student, school and system performance, are major factors that have influenced the development of powerful education management information systems (EMIS).

[4] See: Beaton *et al*. (1996); Mullis *et al*. (1997).

2.0 Essential features of useful PIs

A *useful* performance indicator (PI) is one that informs the processes of strategic decision-making and taking – resulting in measurable improvements to desired outcomes following implementation. Similarly, the *quality* of a PI is comprised of many components including:

- Validity, reliability and relevance to policy;
- Potential for disaggregation (e.g., by gender, socio-economic, ethnic and socio-cultural groupings, education administrations, etc.);
- Timeliness (i.e., currency and punctuality);
- Coherence across different sources;
- Clarity and transparency with respect to known limitations;
- Accessibility and affordability (i.e., cost effectiveness);
- Comparability through adherence to national and internationally agreed standards;
- Consistency over time and location; and
- Efficiency in the use of resources.

The optimum combination of these components is dependent on the uses to be made of the data, since data acceptable for one purpose might be inadequate for another. Thus, because data may be used for many different purposes, the process of determining 'fitness for purpose' is extremely complex and requires wide consultation. The features of five of these characteristics of *useful* PIs, are outlined in more detail below.

Relevance. Judgments related to the *relevance* of a given PI depend on the *purposes* for which it is gathered and *how* it is used to inform policy, planning, practice and reform. Moreover, the *relevance* of any PI is location-specific and context-dependent in terms of prevailing policy priorities and demands for information. In general, however, a PI is deemed to be *relevant* if it provides *useful* information for strategic decision-making and decision-taking. For example, a key guiding principle of the *UNESCO Institute for Statistics* (UIS) in their work of supporting PI data-gathering in 189 member countries and states is that PI data **should not** be collected for their own sake, but rather, because they are needed for specific policy purposes. In this regard, a visit to the UIS website is helpful, at: http://www.unescostat.unesco.org/

Cost-effectiveness. Regardless of the perceived *usefulness* of particular indicators, cost-effectiveness and logistic feasibility are important considerations that need to be taken into account. In the case of indicators of students' achievement outcomes, for example, the cost and feasibility of obtaining estimates derived from full cohort or population data collections may be unjustifiably great compared with those obtained from appropriately designed samples. Decisions about the cost-effectiveness of PIs, however, must be balanced against considerations of their utility to inform policy, planning and reform.

Timeliness. This feature has two key components: *punctuality* and *currency*. Indeed, an important characteristic of the usefulness of PIs is their availability at times when key policy and planning decisions need to be made. At such times, the absence of timely PI information often leads to misinformed enterprises that have a tendency to rely on anecdotal 'myth' and/or opinion rather than on data-informed evidence. Whereas the relevant information for some PIs requires longer periods to collect and analyse (e.g., student achievement progress rates), findings at key stages of the data collection should be reported to inform policy makers and planners of possible trends and other PI factors affecting those trends.

Reliability. Determining the *reliability* of a PI involves evaluating how *accurately* it has been *measured*. This is a crucial technical issue for the formulation and interpretation of PI information that is frequently overlooked by gatherers, purveyors and consumers. Rather, obtaining and reporting evidence concerning the reliability and sources of measurement error for PIs are fundamental responsibilities of PI developers. The same applies to large-scale monitoring procedures employed in national or system-wide testing and public examination

systems that involve the estimation of composite scores from multiple modes of assessment. At the very least, evidence about the uncertainty associated with observed scores is required to minimize the potential 'risks' of misinterpretation.

Validity. This refers to the important issue of *data integrity*. In the present context, however, it should be noted that an estimate of the *reliability* of a PI is not necessarily commensurate with its *validity* – both *content validity* and *criterion-related validity*.[5] While it is possible to have a highly reliable PI that lacks validity (e.g., an assessment task), a *valid* PI that has low reliability is of little or no value. For example, conclusions about students' achievements are *valid* only when measured *reliably* and based on evidence about intended and achieved learning outcomes. Nonetheless, the c*ontent validity* of an indicator – including its *face validity* and *logical validity* (see footnote 5) – may only be established via a rational analysis of its content and utility, based on subjective judgment, albeit by consensus.

In sum, *useful* PIs are those that are *relevant*, *cost-effective*, *timely*, *reliable* and *valid* – in terms of their 'integrity' and capacity to inform the processes of strategic decision-making and decision-taking – resulting in measurable improvements to desired outcomes.

3.0 The 'dangers' of analysing, interpreting and reporting aggregated PI data

More than half a century ago Robinson (1950), writing in the *American Sociological Review*, warned fellow social researchers of the dangers in fitting explanatory regression-type models to aggregated PI data (of the kind relevant here) via the *general linear model* (GLM) under ordinary-least-squares estimation.[i] The fact that many researchers continued to ignore Robinson's warnings, led Cronbach and Webb (1975) to write their paper, which was also ignored (mostly through 'ignorance') by far too many PI data analysts. In response to this 'ignorance', Aitkin and colleagues (Aitkin, Anderson & Hinde, 1981, Aitkin & Longford, 1986), followed by many others,[6] have written extensively on the dangers of fitting explanatory GLM models to aggregated data – in the absence of ALSO fitting the individual-level data from which the aggregated data derive. Without individual-level data, regression analyses of aggregated data results in the well-known phenomenon of ***aggregation bias***[7] – the avoidance of which is ONLY possible by fitting multilevel models to the inherent hierarchically structured data. But what is *aggregation bias*?

Aggregation bias occurs when a variable takes on different meanings and therefore may have different effects at different levels. In educational contexts, for example, a measure of family socio-economic status (SES) aggregated to the school-level from individual student-level data, is often used as measure of a school's or community's resources and normative environment. Whereas the average SES of a school and/or neighbourhood (and their variances) often have

---

[5] ***Content validity*** is established through a rational analysis of the content of an indicator or set of indicators – based on individual, subjective judgment. There are two major types of content validity: face validity and logical validity. ***Face validity*** is established when it is agreed (by consensus) that that an indicator (e.g., a math test score) is a valid measure of a relevant trait (i.e., math achievement). ***Logical*** or ***sampling validity*** involves a careful definition of the **domain** of elements to be measured and the logical design of indicators to cover all the relevant areas of this domain.

Criterion-related validity is established when indicator measures can be related to a predicted criterion. For example, in order to have criterion-related validity, measures of 'inputs' (e.g., per capita cost of education), must be related to (or positively correlated with) a relevant 'output' criterion (student achievement).

[6] For example, see: Goldstein (1986, 1987, 1995, 1997, 2003); Bryk and Raudenbush (1992); Marks, Rowe and Beavis (2003); Raudenbush and Bryk (1988); Raudenbush and Willms (1991); Rowe (1989, 2002a, 2004); and Rowe, Cresswell and Hodgen (2003).

[7] *Aggregation bias* leads to what is commonly known as the "ecological fallacy", namely, relationships between influences and outcomes at the individual-level cannot be derived legitimately from the aggregation of these individual-level data units into groups such as 'schools', 'organisations' or 'areas'.

effects on students' achievements above and beyond the effect of an individual student's SES, the aggregated measure (in the absence of SES and achievement measures at the student-level) becomes a *compressed variance* (or *limited information*) proxy that yields inflated parameter estimates when fitted in regression models (e.g., Figure 3.3 below). Multilevel analysis resolves the confounding of these two effects by facilitating a decomposition of any observed relationship among variables into separate within- and between-school components that are critical to correct interpretations of relationships.

A major contributing factor to *aggregation bias* is the underlying assumption of explanatory GLM-type regression models (see Note 1) that both the fitted response (dependent) and explanatory (independent) variables are **measured without error**! In the case of student achievement measures, and particularly composite measures of SES, this is NOT the case. In this context, Goldstein (1995, p. 8) notes:

> It is well known that when variables in statistical models contain relatively large components of such error the resulting statistical inferences can be very misleading unless careful adjustments are made (Fuller, 1987).

This phenomenon is **compounded** with 'contextual' or 'compositional' variables that are aggregated from the characteristics of level-1 units (i) within level-2 units (j) – or higher – because the measurement error inherent in the level-1 variables is averaged across the level 1 units in each level-2 and higher-level unit. Moreover, there is additional sampling error whenever $n_j < N_i$ – which is always the case.[8]

Nonetheless, the **major** problem associated with fitting aggregated variables in regression models (without also fitting their level-1 derivatives) is the resulting serious compression of variance and consequent loss of information, since we are analysing mere point-estimates with significantly reduced variance (see Figures 1.1 and 1.2 below). Given that the key rationale for fitting regression models to PI data is to ***explain variance*** in the dependent variable(s) as a function of the independent variable(s) (see Note 1), to use aggregated variables only in such models leads to political and ideological-fed delusions of rampant *social determinism*. At this point, a recent example illustrates several key points worth noting.

3.1 Example

As part of the *OECD Programme for International Student Assessment* (PISA) in 2000,[9] the illustrative data presented here consist of *Reading Literacy* achievement scores[10] obtained from a national, stratified, probability-proportional-to-size (PPS) sample of 5176 15-year-old students drawn from 231 secondary schools in 8 Australian states and territories. *Inter alia*, family socio-economic status (SES) was computed as a weighted composite of mother's or father's educational level (whichever was the highest of the two), mother's or father's occupational status (whichever was the highest of the two, and transformed to ANU3), and a measure of family wealth (i.e., the number of specified items in the home). Figure 3.1 provides comparative frequency histograms (on the same metric) of the distributions for students' *Reading Literacy* scores (to the left) and school average scores (to the right), and Figure 3.2 provides similar graphical representations for the SES measure.

---

[8] Note that Fuller (1987) provides a comprehensive account of methods for dealing with measurement errors in linear models, and Goldstein (1995, chp. 10) extends some of those procedures to the multilevel modeling case.

[9] For the first phase of the PISA project during 2000, measures of *Reading Literacy* were obtained from an international sample of 174,896 15 year-old students, drawn from 6638 schools in 32 countries (28 OECD and 4 non-OECD). For published international and Australia-specific findings see: PISA-OECD (2001), and Lokan, Greenwood and Cresswell (2001), respectively.

[10] It is important to note that these 'scores' are transformed logit scores obtained from fitting the item-response data to a Rasch measurement model – one that meets the basic requirements of *valid* measurement (see: Embretson & Hershberger, 1999; Masters, 2001a,b; Masters & Keeves, 1999; Rowe, 1989, 2002a, 2004).

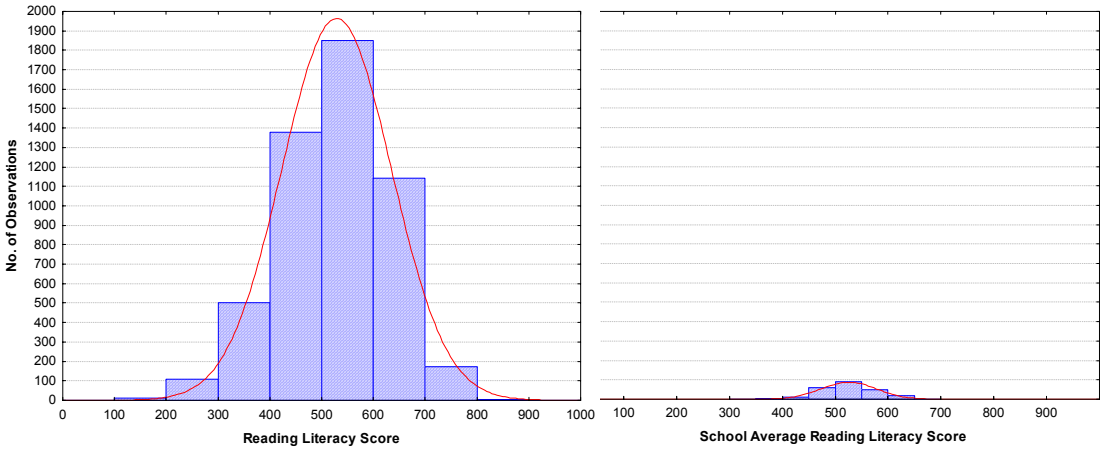

Figure 3.1 Frequency distributions for *Reading Literacy* scores at the individual-level ($Ni = 5176$) **and aggregated average scores at the school-level** ($nj = 231$)

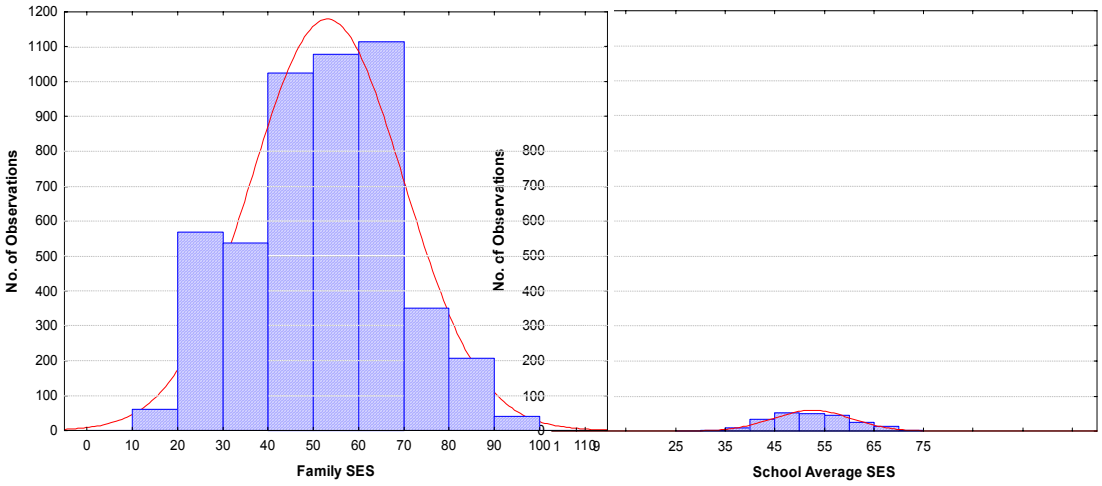

Figure 3.2 Frequency distributions for SES scores at the individual-level ($Ni = 5176$) **and aggregated SES scores at the school-level** ($nj = 231$)

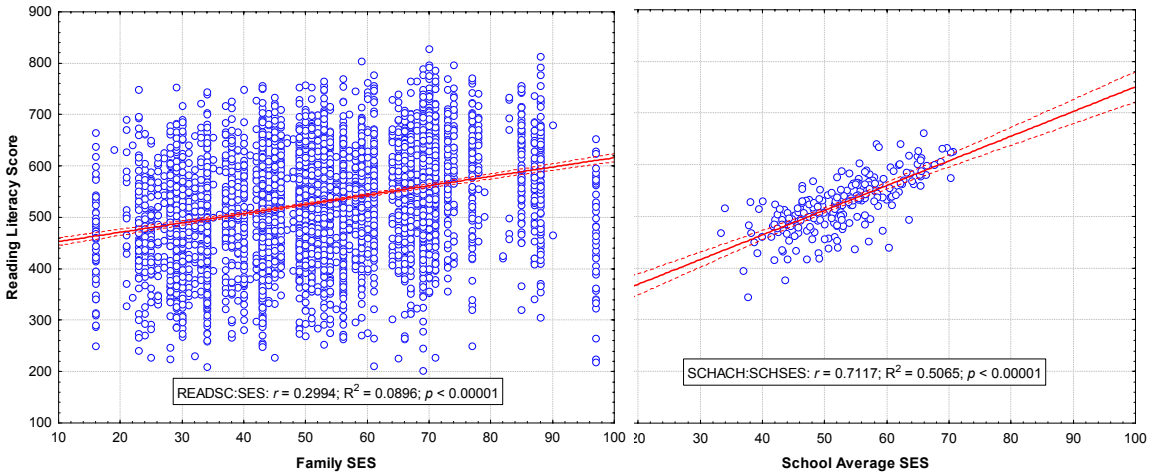

Figure 3.3 Scatter plots illustrating the relationship between *Reading* achievement scores and socioeconomic status (SES) scores at the individual-level ($N_i = 5176$) **and at the school average-level** ($n_j = 231$)**, showing lines of 'best-fit'**

The data summarized in both Figures 3.1 and 3.2 are illustrative of the extreme variance compression (X-axes) that is typically obtained for aggregated data (i.e., from 11054 to 2728 units for *Reading Literacy* scores, and from 284 units to 61 for SES). This compression is also evident in the bivariate scatter plots for *Reading Literacy* achievement and SES given in Figure 3.3. In this latter case, however, the difference in the magnitudes of the regression slopes between the individual-level data ($r_i = 0.299$; $R_i^2 = 0.09$) and aggregated data ($r_j = 0.712$; $R_j^2 = 0.51$) is **massive**. Despite the fact that naïve and ideologically-driven social researchers would be delighted with this 'outcome', such conflations are due to measurement error, variance compression and 'ecological fallacy' pathologies which lead to misrepresentations that only the unwitting and ignorant get excited about – especially economists and sociologists. Moreover, the graphs to the right of Figures 3.1, 3.2 and 3.3 are classic examples of the "ecological fallacy", namely, relationships between influences and outcomes at the individual-level cannot be derived from the aggregation of these individuals into groups such as schools.

A further problem in analysing such data (particularly via linear regression models) with the aim of explaining variance, is violation of the required assumption of Normality. Whereas, many researchers make some reference to the fact that they have 'checked for Normality' – typically via an analysis of residuals – this is almost always done at the univariate level only, NOT at the multivariate level. That is, it is important to examine both the univariate **and** multivariate distributional properties of the continuous variables to be used in subsequent explanatory modelling. For such purposes, PRELIS 2 (Jöreskog & Sörbom, 2003) gives a detailed summary of the distributional parameters (i.e., first-, second-, third- and fourth-order moments) and provides both univariate and multivariate tests of zero skewness and zero kurtosis. To illustrate the importance of checking for both univariate and multivariate estimates of the third- and fourth-order moments (i.e., *skewness* and *kurtosis*, respectively) of the aggregated variables for the above data, the print-out from PRELIS shown below is instructive.

Univariate Summary Statistics for Aggregated Continuous Variables

Variable	Mean	SD .	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
SCHAVACH	523.062	52.227	152.219	-0.080	0.294	344.520	1	661.350	1
SCHAVSES	52.293	7.782	102.137	0.180	-0.464	33.318	1	70.667	1

Tests of Univariate Normality for Aggregated Continuous Variables

	Skewness		Kurtosis		Skewness and Kurtosis	
Variable	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
SCHAVACH	-0.507	0.612	0.981	0.327	1.219	0.544
SCHAVSES	1.135	0.256	-1.780	0.075	4.457	0.108

Tests of Multivariate Normality for Aggregated Continuous Variables

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
0.471	3.015	0.003	8.506	1.1063	0.027	10.303	0.006

Whereas the univariate estimates for skewness and kurtosis are not significantly different from zero, the multivariate estimates are! This result requires that when fitting explanatory regression models to these data, we must normalize the raw data – preferably as normal-equivalent-deviates (NEDs) under the Normal distribution – NOT via log or square-root-type transformations to 'linearize' the data.

4.0 Effective methodologies for the responsible analysis and reporting of PI data: How should such data be analysed?

Before fitting any univariate or multivariate explanatory models to such data, it is vital that key characteristics of the data be examined carefully, namely, their *measurement*, *distributional* and *structural* properties, and the extent to which the analytic and modelling procedures adopted are consistent with clearly articulated substantive research/evaluation questions. In respect of *measurement* properties, it is **vital** that *measurement error* is minimised and accounted for –

otherwise analysts have serious 'garbage-in-garbage-out' problems (see: Embretson & Hershberger, 1999; Masters, 2001a,b; Masters & Keeves, 1999; Rowe, 1989, 2002c, 2004). Whereas the *distributional* properties of such data can be determined (as illustrated above), it is essential that their inherent *structural* properties be taken into account. That is, since the structure of the present data is **hierarchical**, with 5176 students (level-1) grouped within 231 schools (level-2) clustered within 8 Australian States and Territories (level-3), it is vital that this structure is accounted for by fitting ***multilevel models*** to the data. In brief, the rationale for fitting multilevel models to such data is to minimise the risk of: (1) parameter mis-estimation, (2) the likelihood of generating Type I errors due to violation of the assumptions of *independence*, and (3) making erroneous judgements related to statistical conclusion validity (see: Goldstein, 1987, 1995, 2003; Rowe, 1989, 1992, 2002b, 2004; Rowe *et al*., 1995).

4.1 Fitting a baseline variance-components model to the data

In the present case, the basic 3-level variance-components (VC) model for *Reading Literacy* achievement (i.e., within and between-students within schools and States) can be written as:

$$y_{ijk}(\text{readsc}) = \beta_{0ijk} + v_{0k} + u_{0jk} + e_{0ijk} \qquad [4.1]$$

where y_{ikj} (readsc – the response variable of interest) is the transformed logit ***Reading Literacy*** achievement score for student (i) in school (j) and State (k), β_{0ijk} is the 'intercept' term (or grand mean of y_{ijk}), and σ^2_{v0}, σ^2_{u0} and σ^2_{e0} are the residual variances to be estimated for the random terms at the State (v_{0k}), school (u_{0jk}) and student levels (e_{0ijk}), respectively.

Results from the fitted VC model under an *interative generalized least-squares* method of estimation (*IGLS*) via *MLwiN* (Rashbash *et al*., 2003) are given below.[11]

$$\text{readsc}_{studid,\,schlid,\,state} = \beta_{0studid,\,schlid,\,state}\text{cons}$$

$$\beta_{0studid,\,schlid,\,state} = 522.574(6.079) + v_{0state} + u_{0schlid,\,state} + e_{0studid,\,schlid,\,state}$$

$$\left[v_{0state}\right] \sim \text{N}(0,\ \Omega_v) \ :\ \Omega_v = \left[205.814(147.399)\right] \quad \text{Residual variance at the State-level}$$

$$\left[u_{0schlid,\,state}\right] \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \left[2061.521(234.223)\right] \quad \text{Residual variance at the school-level}$$

$$\left[e_{0studid,\,schlid,\,state}\right] \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \left[8824.538(177.458)\right] \quad \text{Residual variance at the student-level}$$

$$\textit{-2*loglikelihood(IGLS Deviance)} = 62142.700(5176 \text{ of } 5176 \text{ cases in use})$$

The grand mean for 15 year-old Australian students' *Reading Literacy* achievement is 522.6 score units (cf. 500 OECD average). Of the 11091.8 total units of residual variation in students' achievement scores (i.e., 205.8 + 2061.5 + 8824.5), an insignificant 1.9% is due to variation between States, but a significant 18.5% is due to between-school differences, and 79.6% at the student-level. The within- and between-State and school residual variances are illustrated in Figure 4.1 below.

To assist interpretation, each horizontal line in Figure 4.1 represents either a State (to the left; $N_{State} = 8$) or a school (to the right; $N_{Schools} = 231$).[12] The length of a line for a State or

[11] Note that ***parameter estimates*** are followed by their ***standard errors*** in parentheses. Statistical significance (at or beyound the $p < 0.05$ α level) is a function of sample size and is indicated when the magnitude of a parameter estimate is at least twice its standard error (i.e., t-value ≥ 1.96 – the univariate 2-tailed 'critial value' under the Normal distribution).

[12] Note that these 'lines' are horizontal since the they derive from the 3-level variance components model described by equation [4.1] which, apart from specifying variation around the grand mean due

school indicates the lowest achievement score in that State or school (at the extreme left of each plot) to the highest score (at the extreme right of each plot). As expected, the largest proportion of variation in students' achievement scores is between-students within-schools (i.e., 79.6%). Nonetheless, there is significant variation between schools' average scores (i.e., 18.5%), ranging from 334 score units to 661 units. Note that it is important not to over-interpret these between-State and between-school estimates since they have not been adjusted for students' intake characteristics (e.g., gender, SES, etc.).

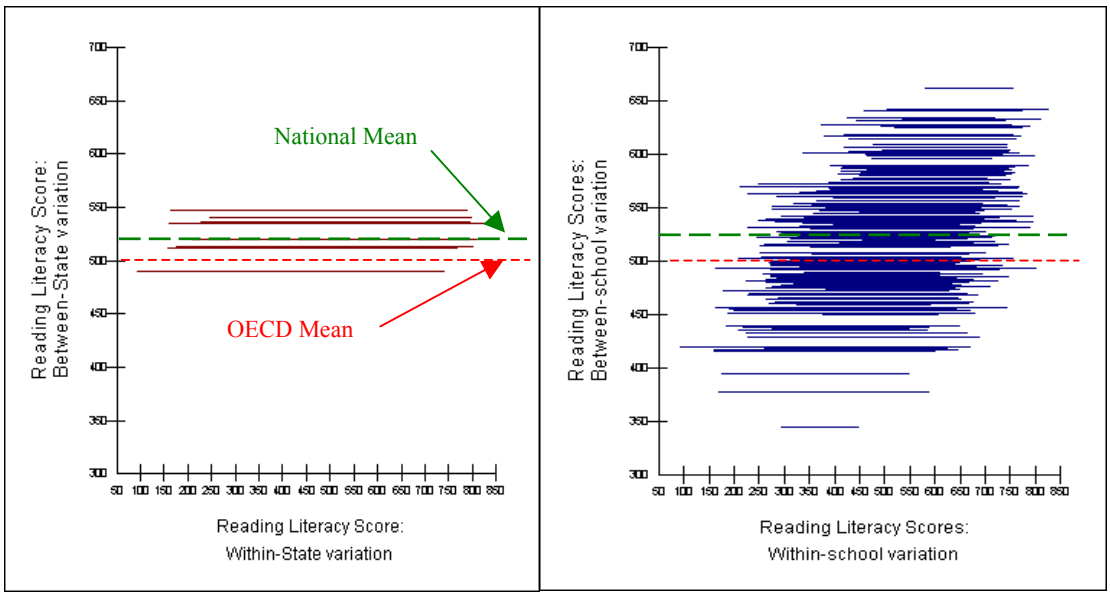

Figure 4.1 Within- and between- State and school variation for *Reading Literacy* scores

To further illustrate the 'danger' of over-interpreting these results, Figure 4.2 provides a type of 'league table' plot of ranked, raw residulas at the State-level.

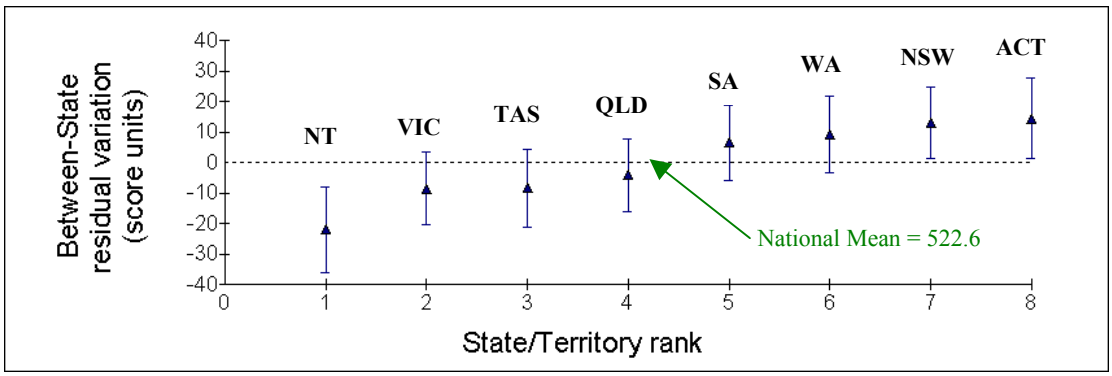

Figure 4.2 Plot of ranked, raw residuals at the State-level, showing mean-point *Reading Literacy* score estimates bounded by 95% 'uncertainty' intervals

Raw, unadjusted, comparative performances of the kind presented graphically in Figure 4.2 have little utility other than to delude public sector administrators (and politicians) by engendering greater or lesser degress of 'self-satistifaction', 'indifference' or 'despair'. Nonetheless, whereas the estimates obtained from fitting a mere variance-components model to the data are not of particular interest (*per se*), they provide a useful baseline from with which to compare more interesting models.

---

to differences between States, schools and students, there are no fitted explanatory variables, and hence, no slopes.

4.2 Fitting an explanatory, multilevel, regression model to the data

In the following model (and solely for illustrative purposes), adjustments are made for the 'intake' variables of gender and family SES (at the student-level), and school average SES at the school-level (i.e., SCHAVSES – to estimate the within-school average effect of SES – over and above that operating at the individual student-level). This model may be written as:

$$y_{ijk}(\text{readsc}) = \beta_{0ijk} + \beta_1 \boldsymbol{x}_{1ijk} + \beta_2 \boldsymbol{x}_{2ijk} + \beta_3 \boldsymbol{x}_{3jk} + v_{0k} + u_{0jk} + e_{0ijk} \qquad [4.2]$$

where y_{ikj} (readsc – the response variable of interest) is the transformed logit ***Reading Literacy*** achievement score for student (i) in school (j) and State (k), and β_{0ijk} is the adjusted 'intercept' (or grand mean of y_{ijk}) after fitting the three explanatory variables of: student gender (sex) ($\boldsymbol{x}_{1ijk}$), family SES ($\boldsymbol{x}_{2ijk}$) at the student-level, and school average SES (SCHAVSES) at the school-level ($\boldsymbol{x}_{3jk}$). The residual variances σ^2_{v0}, σ^2_{u0} and σ^2_{e0} are to be estimated for the random terms at the State (v_{0k}), school (u_{0jk}) and student levels (e_{0ijk}), respectively.

The results of the fitted model described by equation [4.2] to the normalized data[13] are given below, indicating the magnitude of the parameter estimates for the three fitted variables (in SD units), and their respective standard errors (in parentheses).

$$\text{readsc}_{ijk} = \beta_{0jk}\text{cons} + 0.341(0.026)\text{sex}_{ijk} + 0.189(0.014)\text{ses}_{ijk} + 0.259(0.022)\text{schavses}_{jk}$$

$$\beta_{0jk} = -0.156(0.050) + v_{0k} + u_{0jk} + e_{0ijk}$$

$$\begin{bmatrix} v_{0k} \end{bmatrix} \sim \text{N}(0,\ \Omega_v) \ :\ \Omega_v = \begin{bmatrix} 0.016(0.010) \end{bmatrix}$$ Residual variance at the State-level

$$\begin{bmatrix} u_{0jk} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.056(0.009) \end{bmatrix}$$ Residual variance at the school-level

$$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.720(0.015) \end{bmatrix}$$ Residual variance at the student-level

*-2*loglikelihood(IGLS Deviance)* = **12735.290(4979 of 5176 cases in use)**

These results indicate that: (a) the gender effect (in favour of females), and (b) both SES at the student-level and the aggregated SCHAVSES at the school-level, are significant predictors of students' *Reading Literacy* achievement scores. SES at the student-level accounts for < 9% of the unique variance in students' achievement scores (see Figure 3.3), and SCHAVSES accounts for ~12.1% of the unique variance in scores. Together, all three fitted variables account for a mere 21% of the variance in students' achievement scores, with an insignificant 2% of the residual variance at the State-level, and a small but significant 7.1% of the residual variance due to variation between schools.

4.3 Identifying 'better-or-less-than-expected' achievements at the State and school-levels

To identify those States and schools in which students are achieving at lower or higher than expected levels (i.e., net of the effects of the fitted 'intake' variables), analyses of State- and school-level residuals may be undertaken. The results of such analyses are presented graphically in Figures 4.3 and 4.4.

---

[13] Given that the continuous response and exaxplanatory variables of subsequent interest here are measured in different metrics (i.e., READSC, SES and SCHAVSES), such variables should be recomputed as Normal scores, namely as 'normal equivalent deviates' (NEDs) under the Normal distribution, for two reasons: (a) to ensure that such variables are 'measured' on a common metric, and (b) to the assist in the comparative interpretation of 'effect sizes' of the fitted explanatory variables – expressed in terms of standard deviation units (SDs).

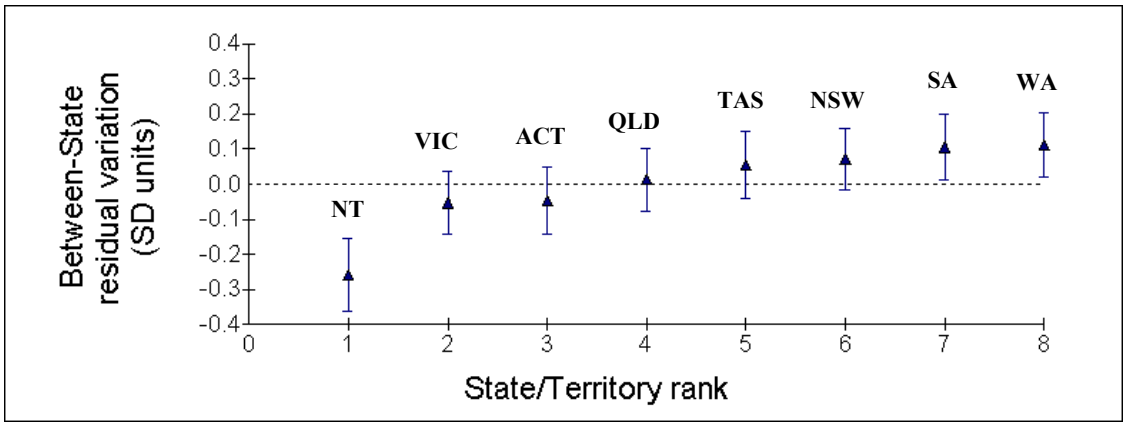

Figure 4.3 Plot of ranked residuals at the State-level, showing adjusted mean-point *Reading Literacy* score estimates bounded by 95% 'uncertainty' intervals[14]

The plot of State-level residuals (Figure 4.3) provides a more 'responsible' representation of between-State comparisons than that shown in Figure 4.2 (p. 10) since, unlike the plot of ranked raw residuals given in Figure 4.2, the estimates shown in Figure 4.3 have at least been adjusted for the 'intake' variables of student gender, family SES and school-average SES.

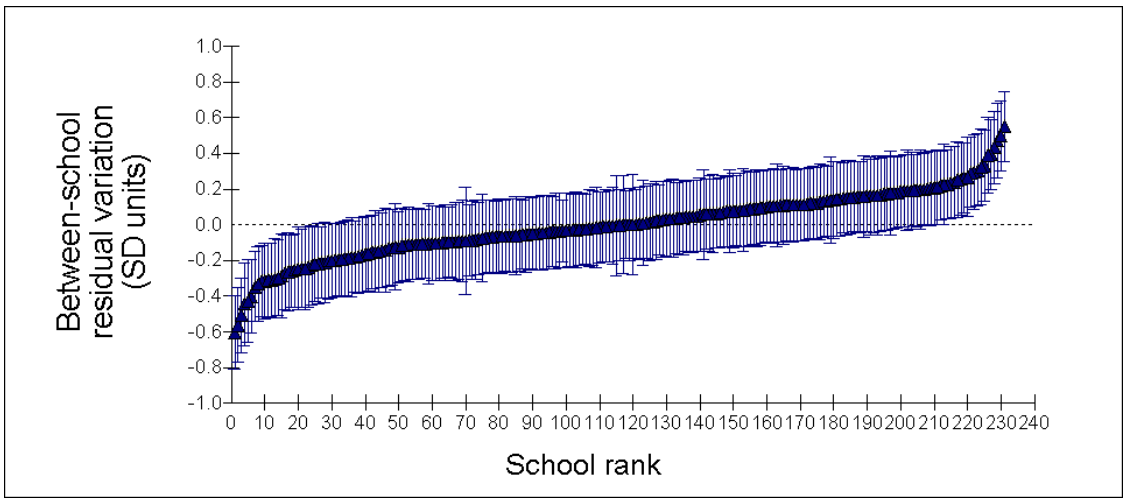

Figure 4.4 Plot of ranked school-level residuals, showing adjusted mean-point *Reading Literacy* score bounded by 95% 'uncertainty' intervals

The plot of school-level residuals given in Figure 4.4 indicates that a few schools have 'intake-adjusted' average achievements significantly **lower** than expected. That is, the 'uncertainty' intervals (see footnote 14) surrounding the adjusted mean-point estimates for these schools are **below** the 'population mean' (zero – indicated by the horizontal dashed line) and do not overlap it. Alternatively, there are a few schools that have 'intake-adjusted' average

[14] Rather than referring to these intervals as ***uncertainty intervals*** (UIs), it is more common to refer to them as ***confidence intervals*** (CIs). Note that 95% *confidence intervals* for a statistic (a mean point-estimate for each school in this case – $\bar{x}_s$) are calculated from: $\bar{x}_s \pm 1.96 \times$ the school's standard error (i.e., the school's standard deviation divided by the square root of the school cohort size – $\sigma_s / \sqrt{n_s}$). These intervals imply that we can be '95% confident' that the estimate of a school's mean lies between these upper and lower limits. However, in the present context of making comparative judgments about the relative performance of schools, these limits are more properly referred to as ***uncertainly intervals***. That is, when the intervals for two or more schools overlap, there is ***no certainty*** that their relative performance differs significantly. For a presentation and discussion of the relevant conceptual and technical issues, see: Goldstein and Healy (1995).

achievements significantly **higher** than expected. In these cases, the 'uncertainty' intervals surrounding the adjusted mean-point estimates for these schools are **above** the 'population' mean, and do not overlap it.

A key feature of Figure 4.4 is the extent to which the uncertainty intervals for each school cover a large part of the total range of estimates, with approximately 84 per cent of the intervals overlapping the population mean (zero). In particular, it illustrates that attempts to separate or rank schools in the form of 'league tables' are subject to considerable uncertainty. Furthermore, there is always the difficulty that any statistical model used to provide such estimates will fail to incorporate **all** the appropriate adjustments (as in the present case), or in some other way may be mis-specified. Thus, at best, even ranked 'value-added' estimates can only be used as screening devices to identify 'outliers' (which could form the basis for follow-up), but they cannot be used as definitive measures of the effect of those schools *per se* on student learning (Rowe, 2000b).

Whereas the use of 'value-added' measures may be able to establish that differences exist among schools or States in the form of 'league table' rankings, they cannot, with any precision, indicate how well a particular school (or State) is performing. The inherent uncertainty of the estimates operates as a fundamental barrier to such knowledge. It should be stressed that raw, or even 'value-added' estimates that are ranked in this way, are ***relative*** ones; that is, they position each school in relation to other schools with which they are being compared, and at a particular point in time. Interpretation of estimates for individual schools is problematic, misleading and potentially irresponsible (see: Goldstein & Spiegelhalter, 1996; Goldstein & Thomas, 1996;). Unfortunately, similar to their counterparts in the UK and USA, Australian politicians and senior bureaucrats currently advocating the publication of such PI 'league tables', are naïvely 'stomping around' in an uninformed epistemopathological fog. Regretfully, this naïvety has been further 'fuelled' by social commentators such as Buckingham (2003).

5.0 Towards responsible analysis, use and reporting of PI data

Throughout industrialised societies there is a prevailing belief that the publication of information about the performance of public bodies is an overwhelming social good. In some societies, such as the United Kingdom and the United States, it is enshrined in public disclosure legislation. In the context of 'public sector effectiveness' the role of published performance information is crucial. Whether intended or not, it provides 'information' for comparative judgements, or in market terms, it introduces a common currency by which the relative 'worth' of public sector service provision is measured. Indeed, this appears to be the primary purpose of such information, and political discourse implicitly acknowledges this whenever reference is made to such matters as 'choice' or 'raising standards'.

As a reaction to unreasonable secrecy, the belief in open access to PI information seems wholly healthy and has led to many benefits. Nevertheless, it can be argued strongly that the public disclosure of information cannot be held to be an *absolute* principle. This is recognised by governments, for example, who normally reserve the right to withhold information deemed to threaten the 'security' of a nation. Similarly, if the publication of certain information has the potential for harming individuals, or may be seriously misleading, then a justifiable case can be mounted for refusing its publication. It could be contended that much of what might be described as *public sector performance indicators* based on 'measures' of client success/ progress (e.g., 'patient throughput' in hospitals, or students' outcomes of schooling) fall into this category. The ability of such 'measures' to reflect objective reality may be extremely limited, and their publication may therefore cause both misleading and incorrect inferences to be drawn about providers and their 'worth' or 'effectiveness'.

In such circumstances, there is strong case for withholding publication. If for whatever reason, publication cannot be prevented then the PI data information should have appropriate warnings attached about its interpretation. By this is meant not simply warnings of the kind that appears on tobacco advertisements, but a proper and prominent *explanation* of why the

information is suspect, together with an assurance that the publishers of the information are fully aware of and accept its limitations.

This view invites criticism of much of the activity that comes under the rubric of *educational performance indicators*. A great deal of this information is produced merely because the data happen to be available. Some of it, such as the achievement scores derived from international studies of mathematics and science (Rotberg, 1990), have been taken, even usurped, by governments and by international agencies such as OECD in order to rank countries in a supposed 'order of merit'. Even where relevant caveats are included in published reports, they tend to be of little avail, and the overall message is that the information presented is useful and informative.

Despite these problems, accountability pressures on governments are not likely to abate in the foreseeable future; nor is the demand for published educational performance indicators based on students' test and examination results obtained from large-scale monitoring programs likely to diminish. Given this 'reality', it is very much in the interests of those wishing to publish such information to consider carefully the need to provide proper guidelines for their publication, if for no other reason than to minimise the risk of widespread public distrust in the face of manifestly poor and misleading information, and to avoid a possible wholesale rejection of all information about schools and schooling – both good and bad. To the writer's credit, Watson (1996, p. 120) recognises the need for such guidelines by proposing three "principles" that "...should underpin any performance indicators framework", namely: (1) the need to develop multiple outcomes, "...which reflect the wide spectrum of objectives for education, not just cognitive outcomes" (ibid.), (2) the need to account for contextualisation factors and to ensure that only 'like-with-like' comparisons are made, and (3) the need for published reports to convey "...the limitations of performance indicators for policy decisions" (ibid.). Watsons' principles constitute a useful start, but given the complexities endemic to the issues involved, a more detailed elaboration is required, particularly for issues related to *publication*.

In an attempt to provide a relevant set of *publication standards*, Goldstein and Myers (1996) propose a set of basic principles for what they refer to as *a code of ethics for performance indicators*; they state:

> Just as educational test constructors have ethical guidelines and in most societies there are codes governing the publication of pornographic or derogatory materials, so we believe there should be a code for the publication of comparative institutional information. ... Our aim is to start a public discussion to see if some consensus can be reached about what a suitable code might contain and whether and how it might be enforced (p. 4).

In promulgating these guidelines, Goldstein and Myers consider the various users of PI information. For example, they suggest that policy makers are interested in broad questions of 'effectiveness' whereas parents and students tend to be more concerned with local details relevant to their particular needs. For all users, however, there is a shared interest in accuracy and general quality and it is these factors which motivate two basic principles:

1. ***The principle of unwarranted harm***. The fundamental guiding principle, as with many ethical codes, is that the publication or communication by other means, should cause no *unwarranted* harm to those who are identified. The term *unwarranted* is used since there will clearly be legitimate circumstances when it is in the 'public interest' for genuinely poor performance to be made known. Nevertheless, the principle is that innocents should be protected from misleading insinuations: for example, implying that a ranking of schools by test or examination scores is also a ranking of educational 'quality' or 'merit'.
2. ***The principle of the right to information***. Given that the information available is believed to accurate and relevant, there shall be a presumption that it be made public, but modified by the first principle where necessary.

These two principles require some elaboration to be applied in practice. The following points can be viewed as offering guidance on the application of principles 1 and 2:

- ***Contextualisation***. PIs should provide information that allow for fair comparisons. Indicators strongly affected by extrinsic/contextual factors (such as student intake characteristics) should not be used unless adjustments have been made for those characteristics. For example, school rankings based solely on 'raw' examination or test score results should not be published. All adjustments for contextual factors should be described carefully and displayed prominently.
- ***Presentation of uncertainty***. All performance indicators should be accompanied by estimates of statistical uncertainty such as those illustrated in Figures 4.3 and 4.4. These should reflect sampling variability, and where possible, the uncertainty due to choice of measurement, statistical techniques used, and so on. The presentation of uncertainty intervals shall be as prominent as those for the indicator values themselves.
- ***Multiple indicators***. Where possible, multiple indicators relevant to each institution should be presented, rather than a single or summary one. This should be done to avoid undue concentration on any one aspect of performance.
- ***Institutional response***. Any institution for which there is a set of published indicators shall have the right to question the accuracy of information about it. Compilers of indicators shall be obliged to make data available in a format which allows for re-analysis of those data by a responsible and competent 'third party', subject to appropriate confidentiality constraints and guided by principle 1.
- ***Agency responsibilities***. Agencies responsible for providing public performance indicators shall assume a responsibility for disseminating accurate and informative material about the underlying procedures used for compilation. They should make relevant technical information accessible, including details of the sampling and statistical methods of analysis used. There is also a responsibility for secondary providers such as the media (newspapers, radio, television) to inform the public of the strengths and limitations of the indicators.
- ***Enforcement***. One would hope that the process of developing such guidelines would generate sufficient awareness of their importance and a common interest in abiding by them. Nevertheless, it may be necessary to establish formal regulatory mechanisms to ensure compliance. This is clearly a matter for careful consideration, but a start in this country might be made with the involvement of professional bodies – independent og government – such as the Australian Council for Educational Research (ACER). Ultimately, the appointment of an educational ombudsperson could provide a means of appropriate redress for aggrieved persons and/or institutions (schools).

In setting out these principles for consideration, the intention of Goldstein and Myers (1996) is to challenge conventional assumptions about the publication of educational performance information, and to highlight the complexity that surrounds these issues. As with any code of ethics, a primary function is to raise awareness of the problems and benefits resulting from particular courses of action. What is important is that persons and institutions should have a means of redress if there is cause to believe they are being unfairly labelled. Moreover, those who are exposed to the PI information should also be exposed to views about its limitations, as well as to its *prima facie* justification. Governments and their bureaucracies have a special responsibility here. Despite prevailing cynicism about officialdom, it is nevertheless the case that the mere fact of publishing PI information by an official body lends it credence. It is therefore important that the publication makes every attempt at honesty and accuracy, since after all, it is the fundamental responsibility of those privileged with access to information and the means to process it, to present it fairly.

6.0 Concluding Comments

Behind the publication of PI information, and especially in the form of 'league tables', lies unspoken assumptions about **data quality** (i.e., coverage/representativeness, and that analyses have taken into account the *measurement*, *distributional* and *structural* properties of the data),

as well as **value judgements** about the location of 'blame' or 'credit'. In the case of educational PIs, the underlying assumption is that if a school is deemed to be 'effective' or 'ineffective' in terms of the ranked position of its students' average test or examination scores on a 'league table', the reason for that performance resides in the school. Even the contextualisation of performance using adjusted or 'value-added' scores may strengthen such an assumption by encouraging the view that **all** other factors have been accounted for, so that any residual variation has its origin in the school. The inherent imprecision of all performance measures and the provisional nature of any conclusions, as argued here, needs to be stressed. Indeed, Saunders (1999, pp. 253-254) expresses a relevant warning in the following terms:

> …both researchers and policy-makers…have a duty to be clear about the fact that there are value judgements as well as conceptual assumptions and technical decisions implicit in what they choose to measure; and that 'value added' measures of effectiveness – powerful as they often are for analytical purposes – are dependent for their credibility on the degree to which those judgements are publicly articulated.

The issue of contextualisation in school education is an important one, but it extends well beyond simplistic notions of 'value-added' indicators of performance. At the very least it needs to be extended to include the general political and social context within which schools operate. Education is not a one-way enterprise. It is not simply the case that the performance of persons in the workplace or society at large can be related causally to their education. To attribute the poor economic performance of a nation to the performance of its education system, for example, is to make both a logical and empirical blunder. It is just as easy to argue the reverse, namely that the economic performance of a nation has direct effects on its education system in terms of motivation, resource provision (e.g., Raffe & Willms, 1991), and other crucial input mechanisms such as the quality and quantity of teachers and their professional development (Hattie, 2003; Hill & Crévola, 1999; Holden, 2004; Rowe, 2000a, 2002c, 2003; Rowe & Hill, 1998; Rowe & Rowe, 1999, 2002). Certainly it is not legitimate to argue, as is frequently the case, that 'league tables' of international educational performance reflect the quality of national educational systems. The attribution of cause and effect is replete with difficulties in such circumstances, and the mere repetition of any given interpretation does not strengthen its plausibility. The same logic applies to the growing use of 'league table'-type PI data to judge the 'quality' or relative 'worth' of individual schools.

The existence of an accountability climate that insists on providing published information that invites comparative judgements about the relative 'worth' of schools – and, inevitably, about the teachers who work in them – is problematic. It is a social and political minefield that has the potential for considerable harm unless it is handled with great care. Again, this is not to deny the usefulness of school-level educational performance indicators involving student achievement data, provided that relevant contextual factors have been taken into account and that the statistical uncertainty associated with the estimates obtained are displayed prominently. McGaw (1991, p. 138) points out the benefits and risks involved in universal achievement monitoring programs in the following terms:

> The benefit of assessing all students is that each school obtains information about its program and teachers obtain potentially helpful diagnostic information about all students. The risk is that the universality of such a program will allow and even encourage comparisons among schools, without consideration of the effect of non-school factors on scores, and so oblige schools to concentrate more upon specific preparation for the tests.

While it would be preferable to implement assessment programs at the beginning of the school year solely for *diagnostic* purposes to assist teachers, as in France (see OECD, 1993), accountability pressures on State and Federal governments in Australia to monitor educational standards are political realities, and ones that are likely to increase. In one sense it could be argued that to propose a control mechanism in the form of a *code of ethics* for the publication of educational performance indicators of the kind outlined above is akin to 'throwing a wet fish at a runaway train'. But if we as a society do nothing, we run the grave risk of rejecting the good and useful information because it cannot be distinguished from the bad and misleading. That, to put it mildly, would be a disaster. An even greater disaster would be, that in our efforts

to meet increasing demands for *assessment, accountability, performance indicators, standards monitoring, quality assurance, school effectiveness* and (now) *benchmarking*, we lose sight of ensuring that what we offer in school education is accessible to **all** students. "The provision of universal education was one of the great social and moral triumphs of the modern period. Universal **success** should be the aim of the post-modern" (Wilson, 1996, p. 8). We stand forever condemned if seduced into diverting the focus of our efforts elsewhere. In the meantime, how should we analyse and report performance indicator data? **'Caress' the data and user beware!**

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society (Series A), 144*, 419-461.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society (Series A), 149*, 1-43.

Alton-Lee, A. (2003). *Quality teaching for diverse students in schooling: Best evidence synthesis*. Wellington, NZ: Medium Term Strategy Policy Division, Ministry of Education [ISBN 0-478-18742-4].

Beaton, A.E, Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Advances in Quantitative Techniques in the Social Sciences, Volume 1. Beverly Hills & London: Sage Publications.

Buckingham, J. (2003). *Schools in the Spotlight: School performance and public accountability* (CIS Policy Monograph 59). Sydney: The Centre for Independent Studies.

Cronbach, L.J., & Webb, N. (1975). Between and within-class effects in a reported aptitude-by-treatment interaction: Reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology*, *6*, 717-724.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6 (1), 1-32.

Embretson, S.E., & Hershberger, S.L. (Eds.) (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fuller, W.A. (1987). *Measurement error models*. New York: Wiley.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Charles Griffin & Co. Ltd.

Goldstein, H. (1995). *Multilevel statistical models* (2nd edn.). London: Edward Arnold.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, *8* (4), 369-395.

Goldstein, H. (2003). *Multilevel statistical models* (3rd edn.). London: Hodder-Arnold.

Goldstein, H., & Healy, M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, A, 158*, 175-177.

Goldstein, H., & Myers, K. (1996). Freedom of information: Towards a code of ethics for performance indicators. University of London Institute of Education.

Goldstein, H., & Spiegelhalter, D. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society, A, 159*, 385-443.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society, A, 159,* 149-163.

Hattie, J. (2003, October). *Teachers make a difference: What is the research evidence?* Background paper to keynote address presented at the 2003 ACER Research Conference, Carlton Crest Hotel, Melbourne, Australia, October 19-21, 2003.

Hill, P.W., & Crévola, C.A. (1999). The role of standards in educational reform for the 21st century. In D.D. Marsh (Ed.), *ASCD Year Book 1999: Preparing our schools for the 21st century* (pp. 117-142). Alexandria, VA: Association for Supervision and Curriculum Development.

Holden, S. (2004). Teachers matter. *Professional Educator, 3* (1), 2-22.

Jöreskog, K.G., & Sörbom, D. (2003). *PRELIS, Version 2.54 for Windows '95/'98/'2000 and Windows NT*. Lincolnwood, IL: Scientific Software International, Inc.

Lokan, J., Greenwood, L., & Cresswell, J. (2001). *The PISA 2000 survey of students' Reading, Mathematical and Scientific Literacy skills: 15-up and counting, reading, writing, reasoning…How literate are Australia's students?* Camberwell, VIC: Australian Council for Educational Research.

Manno, V.B. (1994). *Outcomes-based education: Miracle, cure or plague?* Hudson Institute Briefing Paper No. 165, June 1994.

Marks, G.N., Rowe, K.J., & Beavis, A. (2003). Australian schools not so 'undemocratic'. *Campus Review* (June/July, 2003), p. 34.

Masters, G.N. (2001a). *The key to objective measurement in the psychosocial sciences*. Camberwell, Vic: Australian Council for Educational Research (MIMEO).

Masters, G.N. (2001b). *Educational measurement: ACER Assessment Resource Kit (ARK)*. Camberwell, Vic: Australian Council for Educational Research.

Masters, G.N., & Keeves, J.P. (Eds.) (1999). *Advances in Measurement in Educational Research and Assessment*. New York: Pergamon (Elsevier Science).

McGaw, B. (1991). Monitoring education systems. In J. Chapman, L. Angus, G. Burke, & V. Wilkinson (Eds.) (1991). *Improving the quality of Australian schools.* Australian Education Review No. 33 (pp. 134-139). Hawthorn, Vic: The Australian Council for Educational Research.

Monk, D.H. (1992). Education productivity research: An update and assessment of its role in education finance reform. *Education Evaluation and Policy Analysis, 14*, 307-332.

Mortimore, P. (1991). School effectiveness research: Which way at the crossroads? *School Effectiveness and School Improvement, 2* (3), 213-229.

Mortimore, P. (2001). Globalisation, effectiveness and improvement. *School Effectiveness and Improvement, 12* (2), 229-249.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1997). *Mathematics achievement in the primary school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

NCEE-National Centre on Education and the Economy (1997). *New standards for performance*. Washington, DC:Author.

OECD (1983). *Compulsory schooling in a changing world*. Paris: Organisation for Economic Cooperation and Development.

OECD (1986). *Education and training for manpower development*. Paris: Organisation for Economic Cooperation and Development.

OECD (1989). *Schools and quality: An international report*. Paris: Organisation for Economic Cooperation and Development.

OECD (1993). *Curriculum reform: Assessment in question*. Paris: Organisation for Economic Cooperation and Development.

OECD, (1995). *Indicators of education systems: Measuring the quality of schools*. Paris: Organization for Economic Cooperation and Development.

OECD-PISA. Programme for International Student Assessment (2001). *Knowledge and Skills for Life: First results from PISA 2000*. Paris: Organisation for Economic Co-operation and Development.

Raffe, D, & Willms, J.D. (1991). Schooling the discouraged worker: Local labour-market effects on educational participation. *Sociology, 23*, 559-581.

Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2003). *MLwiN (Beta Version 2): Interactive software for multilevel analysis*. Centre for Multilevel Modelling, Institute of Education, University of London.

Raudenbush, S.W., & Bryk, A.S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E.Z. Rothkopf (Ed.), *Review of Research in Education 1988-1989, Vol. 15* (pp. 423-475). Washington, DC: American Educational Research Association.

Raudenbush, S.W., & Willms, J.D. (Eds.). (1991). *Schools, Classrooms and Pupils: International Studies of Schooling from a Multilevel Perspective*. New York: Academic Press.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351-357.

Rotberg, I.C. (1990). I never promised you first place. *Phi Delta Kappan*, December, 1990: 296-303.

Rowe, K.J. (1989). The commensurability of the general linear model in the context of educational and psychosocial research. *Australian Journal of Education, 33*, 41-52.

Rowe, K.J. (1992). *Identifying Type I errors in educational and social research: Comparisons of results from fitting OLS and multilevel regression models to hierarchically structured data.* Paper presented at the Third National Social Research Conference, The University of Western Sydney, Hawkesbury, June 29 to July 2, 1992.

Rowe, K.J. (2000a). Schooling performance and experiences of males and females: Exploring 'real' effects from evidence-based research in teacher and school effectiveness. *Symposium Proceedings, Educational Attainment and Labour Market Outcomes; Factors Affecting Boys and Their Status in Relation to Girls* (pp. 17-37). Canberra, ACT: Commonwealth Department of Education, Training and Youth Affairs.

Rowe, K. J. (2000b). Assessment, league tables and school effectiveness: Consider the issues and let's get real! *Journal of Educational Enquiry, 1* (1), 72-97.

Rowe, K.J. (2001). Educational performance indicators. In M. Forster, G.N. Masters and K.J. Rowe, Measuring learning outcomes: Options and challenges in evaluation and performance monitoring (pp. 2-20). *Strategic Choices for Educational Reform; Module IV – Evaluation and Performance Monitoring*. Washington, DC: The World Bank Institute.

Rowe, K.J. (2002a, September). *Constructing Performance Indicators for use in Education Management Information Systems (EMISs) for quality system provision: Key features of EMISs leading to teacher and school effectiveness, and to measurable improvements in students' learning and achievement outcomes*. Invited keynote address presented to Uzbekistan Delegation (Monitoring Implementation of Education Reform). Victorian Government Offices, Melbourne, September 17, 2002; available at: http://www.acer.edu.au

Rowe, K.J. (2002b). *The measurement of latent and composite variables from multiple items or indicators: Applications in performance indicator systems*. Background paper to keynote address presented at the RMIT Statistics Seminar Series, October 11, 2002. Camberwell, VIC: Australian Council for Educational Research; available at: http://www.acer.edu.au.

Rowe, K.J. (2002c). The importance of teacher quality. *Issue Analysis*, No. 22, February 27, 2002. Sydney, NSW: Centre for Independent Studies; available at: http://www.cis.org.au

Rowe, K.J. (2003). *The importance of teacher quality as a key determinant of students' experiences and outcomes of schooling*. Background paper to keynote address presented at the 2003 ACER Research Conference, Carlton Crest Hotel, Melbourne, 19-21 October 2003; available at: http://www.acer.edu.au.

Rowe, K.J. (2004). *Practical multilevel analysis with MLwiN & LISREL: An integrated course* (Revised edition). 20th ACSPRI Summer Program in Social Research Methods and Research Technology, Australian National University. Camberwell, VIC: Australian Council for Educational Research.

Rowe, K.J., & Cresswell, J. (2002, September). *Responsible analysis and modelling of PI data: Teacher and school effectiveness*. Invited address and workshop presented for Basic Education Assistance for Mindanao (BEAM): In-Australia Training Program. Camberwell, VIC: Australian Council for Educational Research, September 25, 2002.

Rowe, K.J., Cresswell, J., & Hodgen, E. (2003). *Feasibility of multivariate analyses of school factors relating to achievement:* A research and evaluation report to the Demographic and Statistical Analysis Unit, Data Management and Analysis Division, Ministry of Education, Wellington, New Zealand. Melbourne & Wellington: Australian Council for Educational Research and New Zealand Council for Educational Research**;** available at: http://www.acer.edu.au.

Rowe, K.J., & Hill, P.W. (1998). Modeling educational effectiveness in classrooms: The use of multilevel structural equations to model students' progress. *Educational Research and Evaluation, 4* (4), 307-347.

Rowe, K.J., Hill, P.W, & Holmes-Smith P. (1995). Methodological issues in educational performance and school effectiveness research: A discussion with worked examples (Leading article). *Australian Journal of Education, 39* (3), 217-248.

Rowe, K.J., & Lievesley, D. (2002, April). *Constructing and using educational performance indicators*. Keynote address and workshops presented at the inaugural Asia-Pacific Educational Research Association (APERA) regional conference, ACER, Melbourne, April 16-19, 2002; available at: http://www.acer.edu.au.

Rowe, K.J., & Rowe, K.S. (1999). Investigating the relationship between students' *attentive-inattentive* behaviors in the classroom and their literacy progress. *International Journal of Educational Research*, *31* (1-2 – Whole Issue), 1-138.

Rowe, K.J., & Rowe, K.S. (2002). *What matters most: Evidence-based findings of key factors affecting the educational experiences and outcomes for girls and boys throughout their primary and secondary schooling*. Invited supplementary submission to House of Representatives Standing Committee on Education and Training: *Inquiry Into the Education of Boys* (MIMEO). Melbourne, VIC: Australian Council for Educational Research, and Department of General Paediatrics, Royal Children's Hospital. This submission is available in pdf format at: http://www.acer.edu.au and (submission No. 111) at: http://www.aph.gov.au/house/committee/edt/eofb/index.htm.

Rowe, K.J., Turner, R., & Lane, K. (2002). Performance feedback to schools of students' Year 12 assessments: The *VCE Data Project*. In A.J. Visscher and R. Coe (Eds.), *School improvement through performance feedback* (pp. 163-190). Lisse, the Netherlands: Swetz & Zeitlinger.

Saunders, L. (1999). A brief history of educational 'value added': How did we get to where we are? *School Effectiveness and School Improvement, 10* (2), 233-256.

Summit of the America's (2002). *Line 2*: *Educational Assessment, Brasilia*, March 12-14, 2002. INEP.

Tucker, M.S., & Codding, J.B. (1998). *Standards for our schools: How to set them, measure them and reach them*. San Francisco, CA: Jossey-Bass.

Visscher, A.J., & Coe, R. (2002) (Eds.). *School improvement through performance feedback*. Lisse, the Netherlands: Swetz & Zeitlinger.

Visscher, A., Karsten, S., de Jong, T., & Bosker, R. (2000). Evidence on the intended and unintended effects of publishing school performance indicators. *Evaluation and Research in Education, 14*, 254-267.

Watson, L. (1996). Public accountability or fiscal control? Benchmarks of performance in Australian schooling. *Australian Journal of Education, 40*, 104-123.

Wilson, B. (1996). Current educational priorities, future directions and initiatives. *IARTV Occasional Paper*, No. 45, May, 1996.

Notes

[i] In the context of *explanatory* and *exploratory* approaches to data analysis, some brief notes related to the limitations entailed in omnibus applications of the *general linear model* (GLM) are warranted here. In ***explanatory*** research, it is usually intended to posit some explanation of the relationship between dependent (response) and independent (explanatory) variables, based on *a priori* substantive grounds. Broadly speaking, the explanation is typically formulated in terms of a substantive hypothesis of the kind: "Changes (specified) in the explanatory variables (X's) will give rise to changes (specified) in the dependent variable(s) (Y's)", and typically expressed in a *conditional*, statistical model of the general form:

$$\boldsymbol{Y} = \beta\boldsymbol{X} + \boldsymbol{\varepsilon}, \qquad [A]$$

where $\boldsymbol{Y}$ is a $n \times p$ matrix of observations on p random dependent variables for n cases, $\boldsymbol{X}$ is a known $n \times q$ matrix of observations on q explanatory variables, and β is a $q \times p$ matrix of parameters to be estimated. $\boldsymbol{\varepsilon}$ is a matrix of random prediction residuals whose rows for a given X are uncorrelated, each with mean 0 and common variance-covariance matrix **S**. Specifically, when $\boldsymbol{X}$ is a design matrix (usually 0's and 1's), equation [A] is known as the *general linear model* (GLM). When $\boldsymbol{X}$ represents a matrix of data for observed independent variables on n cases, equation [A] is called the *multivariate regression model* (see Draper & Smith, 1981; Mardia, Kent & Bibby, 1979).

Straightforward extensions to the GLM are employed for multilevel modeling and for covariance structural modeling. It is important to note, however, that both these approaches to explanatory modeling have developed in response to well known limitations in the omnibus use of GLM univariate and multivariate regression-type techniques, including ANOVA, MANOVA, MANCOVA, etc., due to frequent violations of the GLM's underlying assumptions in typical applications. Two major assumptions of the GLM of particular relevance here are: (1) the observed variables are measured without error {i.e., $E(\varepsilon) = 0$; $cov(\varepsilon,\varepsilon') = 0$}; and (2) the models' residuals are 'normally and independently distributed' (NID) with a mean of zero and a variance σ^2 {i.e., $\varepsilon \sim \text{NID}(0, \sigma^2)$}, implying that all variables are measured at a *single-level*, regardless of the structure of the data (for a recent explication of GLM assumptions, see Osborne & Waters, 2002). Nonetheless, for the purpose of highlighting major features of the present distinction, the more familiar GLM will suffice.

Under the above assumptions the observed data are then used in the statistical model (GLM) to test substantive formulations (or hypotheses) using well known least squares, maximum likelihood or Bayesian arguments for obtaining efficient estimates of the parameter coefficients. Strictly speaking, the use of statistical inference is only possible in explanatory research, where the probability statements involved in hypothesis testing are of the form $p(H|D)$; i.e., the "probability of the hypothesis, given the data". The relevant statistical inference involves testing the 'goodness of fit' of the model (characterized by a vector of parameters β) derived from and commensurate with the relevant substantive hypotheses. Moreover, in explanatory research, β should be specified to correspond with the substantive hypotheses and the structure of the data (i.e., single-level or multi-level), as well as the measurement properties of the observed variables. In a strict sense, "...β cannot be specified meaningfully in the absence of substantive hypotheses except in the trivial case of setting all the parameters to zero" (Rowe, 1989, p. 44).

In ***exploratory*** research (including *Data Mining* and *Neural Network* Analysis), however, statistical inference is used to 'scan' the data with the intention of establishing the presence or absence of relations among observations by setting the model's parameters to a null vector. Consequently, for exploratory research (in the absence of substantive hypotheses) statistical significance testing cannot strictly be justified, since the associated question is: "What is the probability of the data, given the hypothesis?" [$p(D|H)$], where the data are approached with an 'open mind'. That is, the research is data-driven, hypothesis-generating, and invariably results in theory conflation. If the parameters of the statistical model are constrained in any way, what little advantages may be gained by exploratory analysis have been destroyed - effectively negating the principle of parsimony.

The exploratory approach is also inadequate in terms of statistical modelling because it does not account for specification in the systematic component of the model ($\boldsymbol{\beta X}$), thus forcing the researcher to specify the relevance of the predictor variables (X's) in a somewhat dubious *post hoc* fashion. It should be noted that a major contribution to the systematic component of a model involves the sampling structure of the observations (i.e., the *representativeness* of the data), thus affecting the estimability of the model. Further, in exploratory research and data analysis, the sufficiency of information to estimate the unknown parameters is often problematic.

In ***explanatory*** research as opposed to ***exploratory*** research, then, the purpose of data analysis is to 'shed light' on substantive theory, but the potential for accomplishing this goal is predicated on the use of statistical models that are commensurate with the substantive model specifications and the characteristics of the data to which they are applied. That is, the legitimacy of 'findings' from fitting such models to data is crucially dependent on taking account of the measurement, distributional and structural properties of those data. A basic understanding of these issues is vital.

Notes References:

Draper, N.R., & Smith, H. (1981). *Applied regression analysis,* (2nd ed.). New York: Academic Press.

Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979). *Multivariate analysis*. New York: Academic Press.

Osborne, J.W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8 (2). This paper is available online at: http://ericae.net/pare/getvn.asp?v=8&n=2.

Rowe, K.J. (1989). The commensurability of the general linear model in the context of educational and psychosocial research. *Australian Journal of Education, 33*, 41-52.